Online adaptive optimal control for continuous-time nonlinear systems with completely unknown dynamics

XXX [a], XXX [a,*], XXX [b], XXX [a] and XXX [a]

a XXX, XXX;
b XXXXX.

(*Received 12 November 2014; accepted XX XX 2014*)

Abstract: An online adaptive optimal control is proposed for continuous-time nonlinear systems with completely unknown dynamics, which is achieved by developing a novel identifier-critic based approximate dynamic programming (ADP) algorithm with a dual neural network (NN) approximation structure. Firstly, an adaptive NN identifier is designed to obviate the requirement of complete knowledge of system dynamics, and a critic NN is employed to approximate the optimal value function. Then the optimal control law is computed based on the information from the identifier NN and the critic NN so that the actor NN is not needed. In particular, a novel adaptive law design method with the parameter estimation error is proposed to online update the weights of both identifier NN and critic NN simultaneously, which converge to small neighborhoods of their ideal values. The closed-loop system stability and the convergence to small vicinity around the optimal solution are all proved by means of the Lyapunov theory. The proposed adaptation algorithm is also improved to achieve finite-time (FT) convergence of the NN weights. Finally, simulation results are provided to exemplify the efficacy of the proposed methods.

Key Words: Adaptive control; optimal control; approximate dynamic programming; system identification; nonlinear systems

1. Introduction

The optimal control is concerned with finding a stabilizing control policy that drives the studied system to a desired target in an optimal way, i.e., to minimize or maximize a predefined performance index or cost function (Lewis, Vrabie, & Syrmos, 2012). Due to its advantages for practical applications, the optimal control has drawn intensive attentions in the control community. Historically, the optimal control problem can be solved by either using the Pontryagin's minimum principle or solving the Hamilton-Jacobi-Bellman (HJB) equation (Lewis et al., 2012; Vrabie & Lewis, 2009). Although mathematically elegant, the HJB equation is generally a nonlinear partial differential equation (PDE), which is intractable to obtain an analytical solution. On the other hand, the dynamic programming (Bellman, 1957) has been used to solve the optimal control problems, which can be implemented backward in time, and thus make the computation to be run with the increased dimension for nonlinear systems. However, these methods are designed in an *offline* manner and require the complete knowledge of system dynamics (Jiang & Jiang, 2012).

Adaptive control (Sastry & Bodson, 1989), on the other hand, has been widely used to investigate the control of systems with unknown parameters and thus to relax the assumptions on the precise system model. However, the associated control action and the error convergence of traditional adaptive control methods are generally not able to minimize the cost function defined in the optimal control framework. Recently, adaptive dynamic programming (ADP) (Werbos, 1992) was proposed to tackle the difficulties encountered in applying dynamic programming to achieve the optimal control, where neural networks (NNs) are used to find the optimal control forward-in-time by using some ideas of adaptive control. In parallel, a bio-inspired method, reinforcement learning (RL) (Doya, 2000; Sutton & Barto, 1998) originally developed in the computational intelligence and machine learning societies has also shown its potential for addressing the optimal control problem. Considering the similarities between ADP and RL, Werbos proposed an actor-critic framework (Werbos, 1990), where neural networks (NNs) are trained to approximate the optimal control based on the named Value Iteration (VI) method. Currently, the ADP-based adaptive optimal control is becoming a cutting-edge research topic (Lewis & Vrabie, 2009; Ni & He, 2013; Qin, Zhang, & Luo, 2014; Si, Barto, Powell, & Wunsch, 2004; F.-Y. Wang, Zhang, & Liu, 2009; Xu, Yang, & Shi, 2014; Zhang, Cui, Zhang, & Luo, 2011).

The ADP schemes have been originally implemented in the iterative manners. Consequently, it is natural to found many successful designs of discrete-time (DT) ADP controls to achieve optimal regulation or tracking (Al-Tamimi, Lewis, & Abu-Khalaf, 2008; Dierks & Jagannathan, 2012; D. Liu & Wei, 2013; Y.-J. Liu, Tang, Tong, Chen, & Li, 2015; D. Wang, Liu, Wei, Zhao, & Jin, 2012; Q. Yang & Jagannathan, 2012; Q. Yang, Vance, & Jagannathan, 2008; Zhang, Song, Wei, & Zhang, 2011). However, extending the ADP control to continuous-time (CT) systems entails challenges in proving the stability and convergence. In fact, early developed ADP algorithms for CT nonlinear systems lack a rigorous stability analysis (Doya, 2000; Hanselmann, Noakes, & Zaknich, 2007). To handle such challenges, an *offline* method has been firstly proposed (Abu-Khalaf & Lewis, 2005) to find an approximate optimal control solution for nonlinear CT systems by incorporating NNs into the actor-critic structure. In the subsequent work (Vrabie & Lewis, 2009; Vrabie, Pastravanu, Abu-Khalaf, & Lewis, 2009), an integral RL technique was designed to get the *online* optimal control based on the Policy Iteration (PI) with a two time-scale actor-critic learning process, i.e., the weights of critic NN and actor NN are updated in a sequential manner (while one NN is tuned the other remains constant). A synchronous ADP algorithm was further proposed in (Vamvoudakis & Lewis, 2010), which uses simultaneous online tuning of actor NN and critic NN by minimizing the Bellman error. A distinct difference between the synchronous ADP and the sequential ADP approaches lies in that both NNs are

* Corresponding author. Email: XXX

trained at the same time in the synchronous ADP (Vamvoudakis & Lewis, 2010). However, these ADP methods are implemented requiring fully known knowledge of system dynamics. Since the exact modeling of nonlinear systems is usually not trivial, it may encounter problem to implement such approaches.

In the control systems, the requirement of system dynamics can be obviated in terms of some observers, e.g., high-gain observers (Farza, Sboui, Cheerier, & M'Saad, 2010) and sliding mode observers (Jung, 2008). In particular, NNs have shown powerful potentials in the observer designs. Inspired by this fact, a novel actor-critic-identifier ADP architecture was proposed (Bhasin et al., 2013), where a NN based identifier is incorporated into the critic-actor framework to estimate the unknown dynamics. A new concurrent learning method was also used in the ADP control (Kamalapurkar, Walters, & Dixon, 2013), where the derivatives of system states are assumed to be measurable. It is noted that the input system dynamics are still required to be known in above ADP control methods; this is slightly stringent in practical applications. To further remove this assumption, recent work (X. Yang, Liu, & Wang, 2014; Zhang, Cui et al., 2011) employed a recurrent neural network (RNN) to fully identify the unknown system dynamics. A similar idea was also used (D. Liu, Huang, Wang, & Wei, 2013) to design an observer based ADP control. Although the states of these identifiers converge to their true values, the convergence of the identifier NN weights cannot be guaranteed (Bhasin et al., 2013; D. Liu et al., 2013; X. Yang et al., 2014; Zhang, Cui et al., 2011). Consequently, only ultimate uniform boundedness (UUB) of the closed-loop system is proved. Only the very recent work (Modares, Lewis, & Naghibi-Sistani, 2013) proposed an adaptive law with the experience replay technique to retain the convergence of the identifier NN weights. In parallel, our previous work (Jing Na & Herrmann, 2014) suggested novel parameter estimation error based adaptive laws for optimal tracking control of unknown nonlinear systems. Nevertheless, it has been found that in the ADP control synthesis (Modares et al., 2013) the convergence of identifier weights is crucial for the convergence of the obtained optimal control. Thus the convergence of identifier or observer should be carefully examined by studying appropriate adaptations in the ADP based optimal control design, in particular for nonlinear CT systems with fully unknown dynamics.

In this paper, we propose a new identifier-critic based ADP algorithm to design optimal control of nonlinear CT systems with completely unknown dynamics, which has a dual approximation structure with an identifier NN and a critic NN. Moreover, novel adaptive laws based on the parameter estimation error (Jing Na, Herrmann, Ren, Mahyuddin, & Barber, 2011) are developed to online update the identifier NN and critic NN weights, such that the convergence of the NN weights to a set around their true values can be proved. This structure is different to aforementioned identifier/observer based ADP methods, e.g., (Bhasin et al., 2013; D. Liu et al., 2013; X. Yang et al., 2014; Zhang, Cui et al., 2011). We first construct an adaptive NN identifier to eliminate the requirement of precisely known system dynamics (including the drift dynamics and input dynamics). Then a critic NN is employed to *online* approximate the solution of the HJB equation. Finally, the estimated optimal value function is used together with the identified dynamics to calculate the optimal control action. Consequently, the widely used actor NN is not needed, which may lead to a simplified identifier-critic ADP structure with dual NN approximators and faster convergence. It should be noted that the proposed '*direct*' parameter estimation scheme based on the parameter estimation error is different to the ideas of minimizing the residual Bellman errors in the HJB equation by using the Least-squares (Bhasin et al., 2013) or the modified Levenberg-Marquardt algorithms (Vamvoudakis & Lewis, 2010). Thus, it is proved that even in the presence of NN approximation errors, the NN weights errors converge to a residual set around zero under a persistent excitation (PE) condition. We also show that the identifier weights error affects the critic NN convergence. The stability of the closed-loop system is proved, and specifically, the convergence of the obtained control to small vicinity around the optimal policy is proved. Finally, the presented adaptations are improved by using the sliding mode technique (Utkin, 1992) to achieve finite-time (FT) convergence for identifier NN and critic NN. Simulation results are given to illustrate the validity of the proposed control schemes.

The contributions can be briefly summarized as: Firstly, a novel identifier-critic based ADP algorithm is proposed to solve the optimal control of nonlinear CT systems. By using online identifier, the assumptions on the unknown dynamics are removed. Moreover, the actor NN is not needed to prove the stability. Thus, instead of the triple-approximation structures (Vamvoudakis & Lewis, 2010; Vrabie et al., 2009; X. Yang et al., 2014; Zhang, Cui et al., 2011), this paper introduces a simplified dual-approximation structure. Secondly, new adaptive laws driven by the parameter estimation error are developed to simultaneously update both the identifier NN and critic NN weights. In particular, these weights are '*directly*' estimated rather than updated to minimize the identifier error and Bellman error. Compared to (Bhasin et al., 2013; Vamvoudakis & Lewis, 2010; Zhang, Cui et al., 2011), the convergence of the identifier NN and critic NN weights to their true values is guaranteed, and thus the proposed control is proved to converge to small vicinity around the optimal solution in this paper.

The paper is organized as follows. The basis of optimal control is given in Section 2. Adaptive optimal control is designed in Section 3. An improved finite-time optimal control is presented in Section 4. Section 5 shows simulations and Section 6 gives the conclusions.

2. Problem formulation

In this paper, we consider the following nonlinear CT system described by

$$\dot{x} = f(x) + g(x)u(t) \tag{1}$$

where $x \in \mathbb{R}^n$ is the measurable system state, $u(t) \in \mathbb{R}^m$ is the control. $f(x) \in \mathbb{R}^n$ is the unknown drift dynamics and $g(x) \in \mathbb{R}^{n\times m}$ is the unknown input dynamics. It is assumed that $f(x)+g(x)u(x)$ is Lipschitz continuous, and system (1) can be stabilized and the solution $x(t)$ is

unique for arbitrary initial state $x(0)$ and control $u(x)$.

The objective of this paper is to design an adaptive control $u(x)$ to stabilize system (1) and to minimize the following infinite-horizon cost function as

$$V(x(t)) = \int_t^\infty r(x(\tau), u(\tau))d\tau \quad (2)$$

where $r(x,u) = x^T Qx + u^T Ru$ is a utility function with Q and R being symmetric positive definite matrices with appropriate dimensions.

Regarding to the optimal control, the designed control $u(x)$ must not only stabilize system (1) but also guarantee that (2) is finite, i.e., the control is admissible (Abu-Khalaf & Lewis, 2005). For this purpose, we define the Hamiltonian of (1) as

$$H(x,u,V) = V_x^T[f(x) + g(x)u] + x^T Qx + u^T Ru \quad (3)$$

where $V_x \triangleq \partial V / \partial x$ denotes the partial derivative of the cost function $V(x)$ with respect to x.

The optimal cost function $V^*(x)$ is given as

$$V^*(x) = \min_{u \in \Psi(\Omega)} \left(\int_t^\infty r(x(\tau), u(x(\tau)))d\tau \right) \quad (4)$$

and it satisfies the HJB equation

$$0 = \min_{u \in \Psi(\Omega)} \left[H(x,u^*,V^*) \right] = V_x^{*T}[f(x) + g(x)u^*] + x^T Qx + u^{*T} Ru^* \quad (5)$$

Then the ideal optimal control u^* can be derived by solving $\partial H(x,u^*,V^*)/\partial u^* = 0$ as

$$u^* = -\frac{1}{2} R^{-1}[g(x)]^T \frac{\partial V^*(x)}{\partial x} \quad (6)$$

where $V^*(x)$ is the solution of the HJB equation (5).

Theoretically, the optimal control for CT nonlinear system (1) can be synthesized from (6). However, this optimal control cannot be obtained from (5) and (6) for practical systems due to the following reasons:

1) In order to calculate control (6), the optimal value function $V^*(x)$ should be obtained by solving the HJB equation (5). However, for general nonlinear systems, the HJB equation is a high-order nonlinear partial differential equation (PDE), which is extremely difficult to be solved by analytical approaches. In particular, the HJB equation is intractable when $f(x), g(x)$ are unknown.

2) The ideal optimal control (6) depends on the input dynamics $g(x)$. Thus the proposed optimal control u^* is not feasible when $g(x)$ is not precisely known.

In order to obtain the optimal control (6), we need to estimate the unknown system dynamics $f(x)$ and $g(x)$ of (1) and derive the solution of the HJB equation (5) using adaptive methods. This has been recently studied by incorporating an identifier into the well-known critic-actor scheme (Bhasin et al., 2013). However, the input dynamics $g(x)$ still needs to be known. Moreover, in the critic-actor algorithm, two NNs (critic NN, actor NN) are used to estimate the value function and the control policy, respectively, such that only UUB of the closed-loop system can be proved.

These issues will be further solved in this paper by introducing a novel identifier-critic ADP architecture (i.e., no actor NN) and new adaptive laws to achieve the convergence of both identifier and optimal control. Thus the fully unknown dynamics $f(x)$ and $g(x)$ can be estimated simultaneously and the suggested dual NN approximation (identifier NN and critic NN) reduces the computational costs.

Definition 1 (Sastry & Bodson, 1989): A vector or matrix function ϕ is persistently excited (PE) if there exist positive constants $\tau > 0, \varepsilon > 0$ such that $\int_t^{t+\tau} \phi(r)\phi^T(r)dr \geq \varepsilon I, \ \forall t \geq 0$.

Lemma 1 (Bhat & Bernstein, 1998): For a CT system $\dot{x} = \phi(x,t), \ \phi(0,t) = 0$, there exist continuously differentiable positive definite function $V(x,t)$ and real numbers $c_1 > 0, 0 < c_2 < 1$ such that $\dot{V}(x,t) \leq -c_1 V^{c_2}(x,t)$ holds, then $V(x,t)$ converges to zero in finite time $t_c \leq \frac{1}{c_1(1-c_2)} V^{1-c_2}(x(t_0),t_0)$ for initial condition $x(t_0)$.

3. Adaptive optimal control design

In this section, an online adaptive ADP algorithm will be proposed to study the optimal control for system (1) by using a NN-based identifier to estimate unknown dynamics and another critic NN to approximate the optimal value function, which are then used to calculate the optimal control action (i.e., the actor NN is avoided). Instead of sequentially updating the critic and actor NNs (Vrabie & Lewis, 2009), both identifier NN and critic NN are updated simultaneously, which leads to an online synchronous learning process. It is noted that the online identifier with classical adaptations may take long time to achieve convergence. To address this issue, a novel method for designing adaptive laws will be suggested to retain fast convergence in this paper. The block diagram of the proposed control system can be shown as in Fig.1.

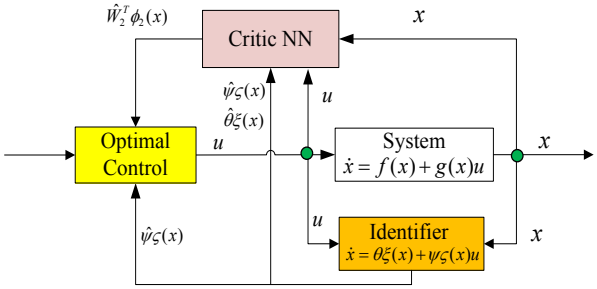

Figure 1. Schematic of the proposed control system.

3.1 Adaptive NN Identifier

An adaptive identifier is firstly constructed to estimate the unknown system dynamics, where NNs are employed. For this purpose, the following assumption is made

Assumption 1 (Abu-Khalaf & Lewis, 2005): The functions $f(x)$, $g(x)$ are continuous on a compact set Ω.

According to Assumption 1, the following linearly parameterized NNs (J. Na, Ren, & Zheng, 2013; Ren, Lewis, & Zhang, 2009) can be used to estimate these unknown functions $f(x)$ and $g(x)$ as

$$f(x) = \theta\xi(x) + \varepsilon_f \quad (7)$$

$$g(x) = \psi\varsigma(x) + \varepsilon_g \quad (8)$$

where $\theta \in \mathbb{R}^{n\times k_\theta}$, $\psi \in \mathbb{R}^{n\times k_\psi}$ are the unknown matrices of NN weights, $\xi \in \mathbb{R}^{k_\theta}$, $\varsigma \in \mathbb{R}^{k_\psi \times m}$ are the basis functions, and $\varepsilon_f \in \mathbb{R}^n$, $\varepsilon_g \in \mathbb{R}^{n\times m}$ are the approximation errors. It is noted that the NN weights θ, ψ and the reconstruction errors ε_f, ε_g are all bounded. Moreover, according to the Weierstrass approximation theorem and the Claims in (Abu-Khalaf & Lewis, 2005; Vamvoudakis & Lewis, 2010), ε_f, ε_g will converge to zero for $k_\theta, k_\psi \to \infty$, which means that the approximation errors vanish as the numbers of NN neuron increase.

From (7) and (8), system (1) can be rewritten as

$$\dot{x} = \theta\xi(x) + \psi\varsigma(x)u + \varepsilon_f + \varepsilon_g u \tag{9}$$

In order to simplify the design of adaptive laws for updating NN weights, system (9) can be represented in a compact form as

$$\dot{x} = W_1^T \phi_1(x,u) + \varepsilon_T \tag{10}$$

where $W_1 = [\theta, \psi]^T \in \mathbb{R}^{(k_\theta + k_\psi)\times n}$ is the augmented unknown weights matrix, $\phi_1(x,u) = [\xi^T(x), u^T\varsigma^T(x)]^T \in \mathbb{R}^{k_\theta + k_\psi}$ is the augmented regressor vector, and $\varepsilon_T = \varepsilon_f + \varepsilon_g u \in \mathbb{R}^n$ denotes the lumped NN error vector.

Remark 1: Several adaptive laws have been proposed to estimate W_1 for system (10), which are designed by minimizing the residual identifier output error (i.e., the error between the system state x and the identifier state $\hat{x}$) based on the gradient (Zhang, Cui et al., 2011) or modified RISE algorithm (Bhasin et al., 2013). In these results, the convergence of the identifier weights to W_1 is not proved though the identifier states $\hat{x}$ converge to their true values x. However, the convergence of the identifier weights W_1 is essential for the convergence of the obtained ADP control (Modares et al., 2013; Jing Na & Herrmann, 2014). Thus this paper will investigate a new adaptive law to guarantee the convergence of the estimation for W_1.

In the following, we will present a novel adaptive law to '*directly*' estimate the unknown NN weights W_1 so that the estimated weights $\hat{W}_1$ converge to the true values W_1. To facilitate further developments, we define the filtered variables x_f, ϕ_{1f} of x, ϕ_1 as

$$\begin{cases} k\dot{x}_f + x_f = x \\ k\dot{\phi}_{1f} + \phi_{1f} = \phi_1 \end{cases} \tag{11}$$

where $k > 0$ is a constant scalar.

For any positive constant $\ell > 0$, we define the filtered regressor matrices $P_1 \in \mathbb{R}^{d\times d}$ and $Q_1 \in \mathbb{R}^{d\times n}$ as

$$\begin{cases} \dot{P}_1 = -\ell P_1 + \phi_{1f}\phi_{1f}^T, & P_1(0) = 0 \\ \dot{Q}_1 = -\ell Q_1 + \phi_{1f}\left[\dfrac{x - x_f}{k}\right]^T, & Q_1(0) = 0 \end{cases} \tag{12}$$

and another auxiliary matrix $M_1 \in \mathbb{R}^{d\times n}$ from P_1, Q_1 as

$$M_1 = P_1\hat{W}_1 - Q_1 \tag{13}$$

Then we can design the adaptive law for $\hat{W}_1$ as

$$\dot{\hat{W}}_1 = -\Gamma_1 M_1 \tag{14}$$

where $\Gamma_1 > 0$ is a constant learning gain matrix.

The parameter k in (11) defines the 'bandwidth' of the filter $(\bullet)_f = (\bullet)/(ks+1)$, which should be set small to retain the robustness. The parameter ℓ in (12) introduces a forgetting factor and also a d.c. gain of $1/\ell$ for the filter $1/(s+\ell)$, thus ℓ should be chosen to tradeoff the convergence speed and the robustness.

Before proving the convergence of adaptive law (14), we present the following Lemmas:

Lemma 2: For variables P_1, Q_1 and M_1 defined in (12) ~(13), then M_1 can be represented as $M_1 = -P_1\tilde{W}_1 + \upsilon_1$, where $\upsilon_1 = -\int_0^t e^{-\ell(t-r)}\phi_{1f}(r)\varepsilon_{Tf}^T(r)dr$ is a bounded variable and $\tilde{W}_1 = W_1 - \hat{W}_1$ is the estimation error.

Proof: For the ordinary matrix differential equation (12), one can obtain its solution as

$$\begin{cases} P_1(t) = \int_0^t e^{-\ell(t-r)}\phi_{1f}(r)\phi_{1f}^T(r)dr \\ Q_1(t) = \int_0^t e^{-\ell(t-r)}\phi_{1f}(r)\left[\dfrac{x(r) - x_f(r)}{k}\right]^T dr \end{cases} \tag{15}$$

On the other hand, it can be verified from (10) and (11) that

$$\dot{x}_f = \frac{x - x_f}{k} = W_1^T\phi_{1f} + \varepsilon_{Tf} \tag{16}$$

where ε_{Tf} is the filtered version of ε_T in terms of a low-pass filter $k\dot{\varepsilon}_{Tf} + \varepsilon_{Tf} = \varepsilon_T$.

Then from (15) and (16), one can obtain that

$$Q_1 = P_1W_1 - \upsilon_1 \tag{17}$$

where $\upsilon_1 = -\int_0^t e^{-\ell(t-r)}\phi_{1f}(r)\varepsilon_{Tf}^T(r)dr$. Considering that the NN basis function $\phi_1(\bullet)$ and error ε_T are all bounded, the variable υ_1 is also bounded, i.e., $\|\upsilon_1\| \le \varepsilon_{\upsilon 1}$ for a positive constant $\varepsilon_{\upsilon 1}$.

Then substituting (17) into (13), the variable M_1 can be rewritten as

$$M_1 = P_1\hat{W}_1 - P_1W_1 + \upsilon_1 = -P_1\tilde{W}_1 + \upsilon_1 \tag{18}$$

It is shown in (18) that the matrix M_1 defined in (13) contains the information of the estimation error $\tilde{W}_1$. Consequently, M_1 can be used to update the NN weights $\hat{W}_1$ as in (14). In particular, the residual error υ_1 will vanish as long as NN approximation error $\varepsilon \to 0$. It is known that $\varepsilon \to 0$ holds for sufficiently large hidden layer NN nodes in the identifier (10), i.e., $d \to +\infty$. □

The positive definite property of matrix P_1 is also crucial for the convergence of $\tilde{W}_1$. Denote $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ as the maximum and minimum eigenvalues of the corresponding matrices, then we have

Lemma 3: If the regressor vector ϕ_1 in (10) is persistently excited (PE), the matrix P_1 defined in (12)

is positive definite, i.e., $\lambda_{\min}(P_1) > \sigma_1 > 0$ holds for a positive constant σ_1.

Proof: We refer to (Jing Na et al., 2011) for the proof. □

The convergence of adaptive law (14) can be given:

Theorem 1: For system (10) with the adaptive law (14), if the regressor vector ϕ_1 is PE, then

i) for $\varepsilon_T = 0$ (i.e., no NN approximation errors), the estimation error $\tilde{W}_1$ converges to zero exponentially.

ii) for $\varepsilon_T \neq 0$ (i.e., with bounded NN approximation errors), the estimation error $\tilde{W}_1$ converges to a compact set around zero.

Proof: We consider the Lyapunov function as $V_1 = \frac{1}{2} tr\left(\tilde{W}_1^T \Gamma_1^{-1} \tilde{W}_1\right)$, then its derivative $\dot{V}_1$ can be calculated by (14) and (18) as

$$\dot{V}_1 = tr\left(\tilde{W}_1^T \Gamma_1^{-1} \dot{\tilde{W}}_1\right) = -tr\left(\tilde{W}_1^T P_1 \tilde{W}_1\right) + tr\left(\tilde{W}_1^T \upsilon_1\right) \quad (19)$$

i) for the case when $\varepsilon_T = 0$, then $\upsilon_1 = 0$ is true, such that (19) can be written as

$$\dot{V}_1 = -tr\left(\tilde{W}_1^T P_1 \tilde{W}_1\right) < -\sigma_1 \left\|\tilde{W}_1\right\|^2 \leq -\mu_1 V_1 \quad (20)$$

where $\mu_1 = 2\sigma_1 / \lambda_{\max}(\Gamma_1^{-1})$ is a positive constant. Then according to the Lyapunov Theorem, the estimation error $\tilde{W}_1$ converges to zero exponentially.

ii) in case $\varepsilon_T \neq 0$, Eq.(19) can be further presented as

$$\dot{V}_1 = -tr\left(\tilde{W}_1^T P_1 \tilde{W}_1\right) + tr\left(\tilde{W}_1^T \upsilon_1\right) \leq -\left\|\tilde{W}_1\right\|(\sigma_1 \left\|\tilde{W}_1\right\| - \varepsilon_{\upsilon 1}) \quad (21)$$

Then according to the extended Lyapunov Theorem, the estimation error $\tilde{W}_1$ uniformly ultimately converges to the compact set $\Omega_1 : \left\{\tilde{W}_1 \mid \left\|\tilde{W}_1\right\| \leq \varepsilon_{\upsilon 1} / \sigma_1\right\}$, of which the size depends on the upper bound of the approximation error $\varepsilon_{\upsilon 1}$ and the excitation level σ_1. □

Remark 2: The condition $\lambda_{\min}(P_1) > \sigma_1 > 0$ is required to prove the convergence of adaptive law (14). Lemma 3 states that this condition can be fulfilled under a conventional PE condition. In general, the online validation of the PE condition is difficult in particular for nonlinear systems. To this end, Lemma 3 also provides a numerically verifiable way to online validate this PE condition, i.e., by calculating the minimum eigenvalue of matrix P_1 to test for $\lambda_{\min}(P_1) > \sigma_1 > 0$. Moreover, the adaptive law (14) is derived without constructing any observer/predictor in comparison to (Bhasin et al., 2013; D. Liu et al., 2013; Modares et al., 2013; X. Yang et al., 2014; Zhang, Cui et al., 2011), and the convergence of the estimation error $\tilde{W}_1$ is guaranteed.

3.2 Adaptive Optimal Control

In this subsection, we propose the optimal control design based on the identified system dynamics. For this purpose, system (1) can be further presented as

$$\dot{x} = \hat{\theta}\xi(x) + \hat{\psi}\varsigma(x)u + \varepsilon_N + \varepsilon_T \quad (22)$$

where $\hat{\theta}$ and $\hat{\psi}$ are the estimations of θ and ψ, respectively, which can be obtained in the estimated matrix $\hat{W}_1$, and $\varepsilon_N = \tilde{W}_1 \phi_1$ is the identifier error. This error will influence the convergence of the proposed control and will be addressed in the stability analysis.

We will find an admissible control $u(x) \in \mu(x)$ such that the cost function (2) associated with system (22) is minimized. For this purpose, the Hamiltonian (3) for system (22) can be rewritten as

$$H(x,u,V) = V_x^T \left[\hat{\theta}\xi(x) + \hat{\psi}\varsigma(x)u + \varepsilon_T + \varepsilon_N\right] + x^T Q x + u^T R u \quad (23)$$

Moreover, the HJB equation (4) becomes

$$\begin{aligned} 0 &= \min_{u \in \Psi(\Omega)} \left[H(x,u^*,V^*)\right] \\ &= V_x^{*T}[\hat{\theta}\xi(x) + \hat{\psi}\varsigma(x)u^* + \varepsilon_T + \varepsilon_N] + x^T Q x + u^{*T} R u^* \end{aligned} \quad (24)$$

Then the optimal control u^* for (22) can be derived by solving HJB equation as

$$u^* = -\frac{1}{2} R^{-1}[\hat{\psi}\varsigma(x)]^T \frac{\partial V^*(x)}{\partial x} \quad (25)$$

where $V^*(x)$ is the solution of the HJB equation (24).

To obtain optimal control (25), one need to solve HJB equation (24) to find the optimal value function $V^*(x)$. However, HJB equation (25) is again a nonlinear PDE. Thus, similar to (Abu-Khalaf & Lewis, 2005; Bhasin et al., 2013; Vamvoudakis & Lewis, 2010; Vrabie & Lewis, 2009; Vrabie et al., 2009), a critic NN will be used to approximate the optimal value function $V^*(x)$. For this purpose, we assume the optimal value function is smooth on the compact set Ω, then there exists a single-layer NN (Vamvoudakis & Lewis, 2010), such that $V^*(x)$ can be uniformly approximated as

$$V^*(x) = W_2^T \phi_2(x) + \varepsilon_v \quad (26)$$

and its derivative is

$$\frac{\partial V^*(x)}{\partial x} = \nabla\phi_2^T W_2 + \nabla\varepsilon_v \quad (27)$$

where $W_2 \in \mathbb{R}^l$ is the ideal weight vector, $\phi_2(x) \in \mathbb{R}^l$ is the basis function vector and ε_v is the approximation error, l is the number of neurons. $\nabla\phi_2 = \partial\phi_2 / \partial x$ and $\nabla\varepsilon_v = \partial\varepsilon_v / \partial x$ are the partial derivative of ϕ_2 and ε_v with respect to x, respectively.

For further study, the following assumption is made

Assumption 2 (Abu-Khalaf & Lewis, 2005): The ideal critic NN weights W_2, the activation function ϕ_2 and its derivative $\nabla\phi_2$ are all bounded, i.e., $\|W_2\| \leq W_N$, $\|\phi_2\| \leq \phi_N$, $\|\nabla\phi_2\| \leq \phi_M$; and the approximation error ε_v and its derivative $\nabla\varepsilon_v$ are bounded, e.g., $\|\nabla\varepsilon_v\| \leq \phi_\varepsilon$.

In practical applications, the NN activation functions $\{\phi_{2i}(e) : i = 1, \cdots, l\}$ can be selected so that $\phi_2(x)$ provides a completely independent basis for V^* as $l \to +\infty$. Then using Assumption 2 and the Weierstrass approximation theorem, both $V^*(x)$ and $\partial V^*(x)/\partial x$ can be uniformly approximated by NNs in (26)~(27), i.e., for $l \to +\infty$, the approximation errors $\varepsilon_v, \nabla\varepsilon_v \to 0$

(Abu-Khalaf & Lewis, 2005; Vamvoudakis & Lewis, 2010).

In the practical control implementation, the critic NN $\hat{V}(x)$ that approximates $V^*(x)$ is given by

$$\hat{V}(x)=\hat{W}_2^T\phi_2(x) \tag{28}$$

where $\hat{W}_2$ is the estimation of the critic NN weights W_2.

From (25) and (28), we get the approximated optimal control u as

$$u=-\frac{1}{2}R^{-1}[\hat{\psi}\varsigma(x)]^T\frac{\partial\hat{V}(x)}{\partial x}=-\frac{1}{2}R^{-1}[\hat{\psi}\varsigma(x)]^T\nabla\phi_2^T(x)\hat{W}_2 \tag{29}$$

where $\partial\hat{V}(x)/\partial x=\nabla\phi_2^T\hat{W}_2$ is the derivative of the critic NN (28) with respect to x.

Remark 3: Available ADP schemes are designed by using another actor NN in conjunction with the critic NN (Abu-Khalaf & Lewis, 2005; Bhasin et al., 2013; Jiang & Jiang, 2012; Vamvoudakis & Lewis, 2010; Vrabie & Lewis, 2009; Vrabie et al., 2009; Zhang, Cui et al., 2011), which may lead to a slightly complicated approximation structure. Moreover, the weights of critic NN and actor NN are updated separately to online minimize the residual Bellman errors in the approximated HJB equation by using the Least-squares (Bhasin et al., 2013) or the modified Levenberg-Marquardt algorithms (Vamvoudakis & Lewis, 2010). However, in the proposed control (29), the critic NN is used to calculate the optimal control action such that the actor NN is avoided; this idea can reduce the computational cost and improve the learning process. Thus the following analysis is different to available ADP schemes.

Now, we will online update the estimated weights $\hat{W}_2$, such that $\hat{W}_2$ converges to a small set around its ideal value W_2. We will extend the idea of Section 3.1 and propose a new estimation scheme based on the Hamiltonian. For this purpose, the approximated HJB equation (24) with critic NN (27) can be rewritten as

$$0=H(x,u,V_x^*)=W_2^T\nabla\phi_2\left(\hat{\theta}\xi+\hat{\psi}\varsigma u\right)+x^TQx+u^TRu+\varepsilon_{HJB} \tag{30}$$

where $\varepsilon_{HJB}=W_2^T\nabla\phi_2(\varepsilon_N+\varepsilon_T)+\nabla\varepsilon_v\left(\hat{\theta}\xi+\hat{\psi}\varsigma u+\varepsilon_T+\varepsilon_N\right)$ is a bounded residual HJB equation error due to the NN approximation errors ε_N, ε_T and $\nabla\varepsilon_v$, which can be made arbitrarily small with sufficiently large NN nodes (Abu-Khalaf & Lewis, 2005; Vamvoudakis & Lewis, 2010), i.e., $\varepsilon_N,\varepsilon_T\to 0$ for $k_\theta,k_\psi\to+\infty$ and $\nabla\varepsilon_v\to 0$ for $l\to+\infty$. It is also shown in (30) that the convergence of the identifier weight error $\tilde{W}_1$ to zero is crucial for the convergence of critic NN because of the induced identifier error $\varepsilon_N=\tilde{W}_1\phi_1$ in ε_{HJB}.

To facilitate the design of adaptive law, we denote the known terms in (30) as $\Xi=\nabla\phi_2\left(\hat{\theta}\xi+\hat{\psi}\varsigma u\right)$ and $\Theta=x^TQx+u^TRu$, so that the approximated HJB equation (30) is given as

$$\Theta=-W_2^T\Xi-\varepsilon_{HJB} \tag{31}$$

As shown in (31), the unknown critic NN weights W_2 appear in a linearly parameterized form, and thus can be '*directly*' estimated by extending the adaptation proposed in Section 3.1. Then we define the filtered regressor matrix $P_2\in\mathbb{R}^{l\times l}$ and vector $Q_2\in\mathbb{R}^l$ as

$$\begin{cases}\dot{P}_2=-\ell P_2+\Xi\Xi^T, & P_2(0)=0\\ \dot{Q}_2=-\ell Q_2+\Xi\Theta, & Q_2(0)=0\end{cases} \tag{32}$$

where $\ell_2>0$ is a constant. Another auxiliary vector $M_2\in\mathbb{R}^l$ is calculated based on P_2 and Q_2 as

$$M_2=P\hat{W}_2+Q_2 \tag{33}$$

Then the adaptive law for the critic NN is designed as

$$\dot{\hat{W}}_2=-\Gamma_2M_2 \tag{34}$$

where $\Gamma_2>0$ is a constant learning gain.

Similar to Lemma 2 and Lemma 3, we have

Lemma 4: For variables P_2, Q_2 and M_2 defined in (32) ~(33), then M_2 can be represented as $M_2=-P_2\tilde{W}_2+\upsilon_2$, where $\upsilon_2=-\int_0^t e^{-\ell(t-r)}\varepsilon_{HJB}(r)\Xi^T(r)dr$ is a bounded variable, i.e., $\|\upsilon_2\|\le\varepsilon_{\upsilon 2}$ for a positive constant $\varepsilon_{\upsilon 2}$ and $\tilde{W}_2=W_2-\hat{W}_2$ is the estimation error.

The proof of Lemma 4 can be conducted by solving the equation (32) with $Q_2=-P_2W_2+\upsilon_2$ and following similar mathematical manipulations to Lemma 2. Note the variable M_2 contains the estimation error $\tilde{W}_2$, and thus can be used to drive the adaptive law (34).

Lemma 5: If the regressor vector Ξ in (31) is PE, then the matrix P_2 defined in (32) is positive definite, i.e., $\lambda_{\min}(P_2)>\sigma_2>0$ for a positive constant $\sigma_2>0$.

We now summarize the results of this subsection as:

Theorem 2: For adaptive law (34) of critic NN with the regressor vector Ξ in (31) being PE, then

i) for $\varepsilon_{HJB}=0$ (i.e., no NN approximation errors), the critic NN error $\tilde{W}_2$ converges to zero exponentially.

ii) ii) for $\varepsilon_{HJB}\neq 0$ (i.e., with NN approximation errors), the critic NN error $\tilde{W}_2$ converges to a compact set around zero.

The proof of Theorem 2 is similar to that of Theorem 1 by considering the adaptive law (34) with Lemma 4 and Lemma 5. The only essential difference is the critic NN weights $\hat{W}_2$ is a vector but not a matrix as the identifier NN weights, thus the Lyapunov function should be selected as $V_2=\frac{1}{2}\tilde{W}_2^T\Gamma_2^{-1}\tilde{W}_2$. Here, the detailed proof will not be provided due to the page limit.

Remark 4: It is shown in (30) that the residual HJB equation error ε_{HJB} is due to the critic NN approximation error $\nabla\varepsilon_v$ in (27), the identifier errors $\varepsilon_N=\tilde{W}_1\phi_1$ and ε_T. Then the convergence of the identifier weights to their true values is essential for the convergence of the critic NN weights, and thus the proposed optimal control action. This issue is fully addressed in this paper by

introducing novel parameter estimation error based adaptive laws (14) and (34), which are clearly different to most of available results, e.g., (Bhasin et al., 2013; X. Yang et al., 2014; Zhang, Cui et al., 2011).

3.3 Stability Analysis

This subsection presents the stability analysis. For this purpose, the system dynamics with the proposed optimal control is first studied. By substituting the optimal control (29) into (1), one have the system dynamics as

$$\dot{x} = f(x) + g(x)\left(-\frac{1}{2}R^{-1}\hat{g}^T(x)\nabla\phi_2^T\hat{W}_2 + \frac{1}{2}R^{-1}g^T(x)(\nabla\phi_2^T W_2 + \nabla\varepsilon_v)\right) + g(x)u^* \tag{35}$$

where $\hat{g}(x) = \hat{\psi}\varsigma$ denote the estimation of the input dynamics $g(x)$, which is given in the identifier (14).

In this case, we can further obtain $g^T\nabla\phi_2^T W_2 - \hat{g}^T\nabla\phi_2^T\hat{W}_2 = g^T\nabla\phi_2^T\tilde{W}_2 + \tilde{g}^T\nabla\phi_2^T\hat{W}_2$, so that (35) can be rewritten as

$$\dot{x} = f(x) + \frac{1}{2}gR^{-1}\left(g^T\nabla\phi_2^T\tilde{W}_2 + \tilde{g}^T\nabla\phi_2^T\hat{W}_2\right) + gu^* + \frac{1}{2}gR^{-1}g^T\nabla\varepsilon_v \tag{36}$$

It should be noted that the dynamics $f(x)$, $g(x)$ are unknown and only the estimated dynamics can be used. In this case, the effect of the estimation error $\varepsilon_N = \tilde{W}_1\phi_1$ on the convergence of the proposed optimal control has to be considered in the Lyapunov function. Consequently, the following stability analysis is different to some available results, e.g., (Vamvoudakis & Lewis, 2010).

To facilitate the stability analysis, the following assumption used in the literature (e.g., (Hanselmann et al., 2007; Modares et al., 2013)) is made:

Assumption 3: The dynamics of system (1) fulfill the condition $\|f(x)\| \le b_f\|x\|, \|g(x)\| \le b_g$ for some positive constants $b_f > 0$, $b_g > 0$.

We now summarize the main results of this paper as:

Theorem 3: For system (1) with adaptive optimal control (29) and adaptive laws (14) and (34), if the regressor vectors ϕ_1 and Ξ are PE, then

i) in the absence of NN approximation errors, the system state x and the NN weights errors $\tilde{W}_1$, $\tilde{W}_2$ converge to zero, and the adaptive control u in (29) converges to the ideal optimal solution u^* in (6), i.e., $u \to u^*$.

ii) in the presence of NN approximation errors, the system state x and the NN weights errors $\tilde{W}_1$, $\tilde{W}_2$ are UUB, and the adaptive control u in (29) converges to a small region around its optimal solution u^* in (6), i.e., $\|u - u^*\| \le \varepsilon_u$ for a positive constant ε_u.

Proof: Consider the Lyapunov function as

$$V = V_1 + V_2 + V_3 + V_4 + V_5 = \frac{1}{2}tr\left(\tilde{W}_1^T\Gamma_1^{-1}\tilde{W}_1\right) + \frac{1}{2}\tilde{W}_2^T\Gamma_2^{-2}\tilde{W}_2 + \Gamma x^T x + KV^* + \Upsilon_1\upsilon_1^T\upsilon_1 + \Upsilon_2\upsilon_2^T\upsilon_2 \tag{37}$$

where V^* is the optimal cost function defined in (4) and $K > 0, \Gamma > 0, \Upsilon_1 > 0, \Upsilon_2 > 0$ are positive constants.

Consider the inequality $ab \le a^2\eta/2 + b^2/2\eta$ with $\eta > 0$, then we can obtain from (14) and (34) that

$$\begin{aligned}\dot{V}_1 &= -tr\left(\tilde{W}_1^T P_1\tilde{W}_1\right) + tr\left(\tilde{W}_1^T\upsilon_1\right) \le -\sigma_1\left\|\tilde{W}_1\right\|^2 + \left\|\tilde{W}_1^T\upsilon_1\right\| \\ &\le -(\sigma_1 - \frac{1}{2\eta})\left\|\tilde{W}_1\right\|^2 + \frac{\eta\|\upsilon_1\|^2}{2}\end{aligned} \tag{38}$$

and

$$\begin{aligned}\dot{V}_2 &= -\tilde{W}_2^T P_2\tilde{W}_2 + \tilde{W}_2^T\upsilon_2 \le -\sigma_2\left\|\tilde{W}_2\right\|^2 + \left\|\tilde{W}_2^T\upsilon_2\right\| \\ &\le -(\sigma_2 - \frac{1}{2\eta})\left\|\tilde{W}_2\right\|^2 + \frac{\eta\|\upsilon_2\|^2}{2}\end{aligned} \tag{39}$$

Moreover, one may get $\dot{V}_3$ from (4) and (35) as

$$\begin{aligned}\dot{V}_3 &= 2\Gamma x^T\dot{x} + K(-x^TQx - u^{*T}Ru^*) \\ &= 2\Gamma x^T\left(f + \frac{1}{2}gR^{-1}\left(g^T\nabla\phi_2^T\tilde{W}_2 + \tilde{g}^T\nabla\phi_2^T\hat{W}_2\right) + gu^* + \frac{1}{2}gR^{-1}g^T\nabla\varepsilon_v\right) + K(-x^TQx - u^{*T}Ru^*) \\ &\le -\left[K\lambda_{\min}(Q) - 2b_f\Gamma - \left(\eta b_g^2\phi_M\lambda_{\max}(R^{-1}) + \eta b_g\phi_M b_w\lambda_{\max}(R^{-1}) + 2\right)\right]\|x\|^2 + \frac{1}{4\eta}\Gamma^2 b_g^2\phi_M\lambda_{\max}(R^{-1})\left\|\tilde{W}_2\right\|^2 \\ &\quad + \frac{1}{4\eta}\Gamma^2 b_g\phi_M b_w\lambda_{\max}(R^{-1})\left\|\tilde{W}_1\right\|^2 + \frac{1}{4}\Gamma^2 b_g^4\lambda_{\max}^2(R^{-1})\nabla\varepsilon_v^T\nabla\varepsilon_v \\ &\quad - \left(K\lambda_{\min}(R) - \Gamma^2 b_g^2\right)\|u^*\|^2\end{aligned} \tag{40}$$

where $b_w = \left\|\hat{W}_2\right\|$ is a bounded variable.

From (17), it is evident that is $\dot{\upsilon}_1 = -\ell\upsilon_1 + \phi_{1f}\varepsilon_{Tf}^T$, so that

$$\begin{aligned}\dot{V}_4 &= 2\Upsilon_1\upsilon_1^T\dot{\upsilon}_1 = 2\Upsilon_1\upsilon_1^T\left(-\ell\upsilon_1 + \phi_{1f}\varepsilon_{Tf}^T\right) \\ &\le -(2\Upsilon_1\ell - \eta)\|\upsilon_1\|^2 + \frac{1}{\eta}\left\|\Upsilon_1\phi_{1f}\varepsilon_{Tf}^T\right\|^2\end{aligned} \tag{41}$$

Moreover, one may obtain from (30) that $\varepsilon_{HJB} = W_2^T\nabla\phi_2(\varepsilon_N + \varepsilon_T) + \nabla\varepsilon_v\left[f(x) + g(x)u\right]$ holds, so that $\dot{\upsilon}_2 = -\ell\upsilon_2 + \Xi\varepsilon_{HJB}$ can be given as

$$\begin{aligned}\dot{V}_5 &= 2\Upsilon_2\upsilon_2^T\upsilon_2 \\ &= 2\Upsilon_2\upsilon_2^T\left\{-\ell\|\upsilon_2\| + \Xi\left[W_2^T\nabla\phi_2(\varepsilon_N + \varepsilon_T) + \nabla\varepsilon_v\left(f + gu\right)\right]\right\} \\ &\le -(2\Upsilon_2\ell - 4\eta)\|\upsilon_2\|^2 + \frac{1}{\eta}\Upsilon_2^2 W_N^2\phi_M^2\|\Xi\|^2\|\varepsilon_N\|^2 \\ &\quad + \frac{1}{\eta}\Upsilon_2^2 W_N^2\phi_M^2\|\Xi\|^2\|\varepsilon_T\|^2 + \frac{1}{\eta}\Upsilon_2^2 b_f^2\phi_\varepsilon^2\|\Xi\|^2\|x\|^2 \\ &\quad + \frac{1}{4\eta}\Upsilon_2^2 b_g^2 b_w^2 b_\psi^2\phi_M^2\lambda_{\max}^2(R^{-1})\|\Xi\|^2\nabla\varepsilon_v^T\nabla\varepsilon_v\end{aligned} \tag{42}$$

where $b_\psi = \left\|\hat{\psi}\varsigma\right\|$ is a bounded variable.

Consequently, we substitute $\varepsilon_N = \tilde{W}_1\phi_1$ into (42) and

thus have

$$\begin{aligned}\dot{V} &= \dot{V}_1 + \dot{V}_2 + \dot{V}_3 + \dot{V}_4 + \dot{V}_5 \\ &\le -\left(\sigma_1 - \frac{1}{2\eta} - \frac{1}{4\eta}\Gamma^2 b_g \phi_M b_w \lambda_{\max}(R^{-1}) - \frac{1}{\eta}\Upsilon_2^2 W_N^2 \phi_M^2 \|\Xi\|^2 \|\phi_1\|^2\right)\|\tilde{W}_1\|^2 \\ &\quad -\left(\sigma_2 - \frac{1}{2\eta} - \frac{1}{4\eta}\Gamma^2 b_g^2 \phi_M \lambda_{\max}(R^{-1})\right)\|\tilde{W}_2\|^2 \\ &\quad -\Big[K\lambda_{\min}(Q) - 2\Gamma b_f - \frac{1}{\eta}\Upsilon_2^2 b_f^2 \phi_\varepsilon^2 \|\Xi\|^2 \\ &\quad -\left(\eta b_g^2 \phi_M \lambda_{\max}(R^{-1}) + \eta b_g \phi_M b_w \lambda_{\max}(R^{-1}) + 2\right)\Big]\|x\|^2 \\ &\quad -\left(2\Upsilon_1 \ell - \frac{3\eta}{2}\right)\|\upsilon_1\|^2 - \left(2\Upsilon_2 \ell - \frac{9\eta}{2}\right)\|\upsilon_2\|^2 - \left(K\lambda_{\min}(R) - \Gamma^2 b_g^2\right)\|u^*\|^2 \\ &\quad +\left(\frac{1}{4}\Gamma^2 b_g^4 \lambda_{\max}^2(R^{-1}) + \frac{1}{4\eta}\Upsilon_2^2 b_g^2 b_w^2 b_\psi^2 \phi_M^2 \lambda_{\max}^2(R^{-1})\|\Xi\|^2\right)\|\nabla \varepsilon_v\|^2 \\ &\quad +\frac{1}{\eta}\|\Upsilon_1 \phi_{1f} \varepsilon_{1f}^T\|^2 + \frac{1}{\eta}\Upsilon_2^2 W_N^2 \phi_M^2 \|\Xi\|^2 \|\varepsilon_T\|^2\end{aligned} \tag{43}$$

Clearly, we can choose the parameters $K, \Gamma, \Upsilon_1, \Upsilon_2, \eta$ fulfilling the following conditions

$$K > (\eta b_g^2 \phi_M \lambda_{\max}(R^{-1}) + \eta b_g \phi_M b_w \lambda_{\max}(R^{-1}) + 2 + 2\Gamma b_f + \Upsilon_2^2 b_f^2 \phi_\varepsilon^2 \|\Xi\|^2 \frac{1}{\eta})/\lambda_{\min}(Q)$$

$$\eta > \max\left\{\left(2 + \Gamma b_g \phi_M b_w \lambda_{\max}(R^{-1}) + 4\Upsilon_2^2 W_N^2 \phi_M^2 \|\Xi\|^2 \|\phi_1\|^2\right)/4\sigma_1,\right.$$
$$\left.\left(2 + \Gamma b_g^2 \phi_M \lambda_{\max}(R^{-1})\right)/4\sigma_2\right\}$$

$$\Gamma > b_g^2/\lambda_{\min}(R), \quad \Upsilon_1 > \frac{3\eta}{4\ell}, \quad \Upsilon_2 > \frac{9\eta}{4\ell}.$$

Then (43) can be further presented as

$$\dot{V} \le -a_1\|\tilde{W}_1\|^2 - a_2\|\tilde{W}_2\|^2 - a_3\|x\|^2 - a_4\|\upsilon_1\|^2 - a_5\|\upsilon_2\|^2 + \gamma \tag{44}$$

where a_1, a_2, a_3, a_4 and a_5 are positive constants defined by

$$a_1 = \sigma_1 - \frac{1}{2\eta} - \frac{1}{4\eta}\Gamma^2 b_g \phi_M b_w \lambda_{\max}(R^{-1}) - \frac{1}{\eta}\Upsilon_2^2 W_N^2 \phi_M^2 \|\Xi\|^2 \|\phi_1\|^2,$$

$$a_2 = \sigma_2 - \frac{1}{2\eta} - \frac{1}{4\eta}\Gamma^2 b_g^2 \phi_M \lambda_{\max}(R^{-1}),$$

$$a_3 = K\lambda_{\min}(Q) - \left(\eta b_g^2 \phi_M \lambda_{\max}(R^{-1}) + \eta b_g \phi_M b_w \lambda_{\max}(R^{-1})\right) - 2 - 2\Gamma b_f - \frac{1}{\eta}\Upsilon_2^2 b_f^2 \phi_\varepsilon^2 \|\Xi\|^2,$$

$$a_4 = 2\Upsilon_1 \ell - 3\eta/2,$$

$$a_5 = 2\Upsilon_2 \ell - 9\eta/2,$$

$$\gamma = \left(\frac{1}{4}\Gamma^2 b_g^4 \lambda_{\max}^2(R^{-1}) + \frac{1}{4\eta}\Upsilon_2^2 b_g^2 b_w^2 b_\psi^2 \phi_M^2 \lambda_{\max}^2(R^{-1})\|\Xi\|^2\right)\|\nabla \varepsilon_v\|^2 + \frac{1}{\eta}\|\Upsilon_1 \phi_{1f} \varepsilon_{1f}^T\|^2 + \frac{1}{\eta}\Upsilon_2^2 W_N^2 \phi_M^2 \|\Xi\|^2 \|\varepsilon_T\|^2.$$

1) In case when there are no approximation errors in both identifier NN and critic NN, i.e., $\varepsilon_T = \nabla \varepsilon_v = 0$ and thus $\gamma = 0$, then (44) can be rewritten as

$$\dot{V} = -a_1\|\tilde{W}_1\|^2 - a_2\|\tilde{W}_2\|^2 - a_3\|x\|^2 - a_4\|\upsilon_2\|^2 \le 0 \tag{45}$$

Thus, according to the Lyapunov Theorem, $V \to 0$ holds for $t \to +\infty$, such that the estimation error $\tilde{W}_1$, $\tilde{W}_2$ and the system states x all converge to zero.

Moreover, in this case by assuming $\varepsilon_g = 0$, we know $\hat{W}_1 \to W_1$ and $\hat{W}_2 \to W_2$ so that $\hat{\psi}\hat{\varsigma}(x) \to g(x)$ holds for $t \to +\infty$. Thus it can be obtained that the error between the ideal optimal control u^* in (6) and the proposed approximated optimal control u in (29) can be represented as

$$\begin{aligned}u - u^* &= \frac{1}{2}R^{-1}\left(g^T \nabla \phi_2^T \tilde{W}_2\right) + \frac{1}{2}R^{-1}[g - \hat{\psi}\hat{\varsigma}]^T \nabla \phi_2^T W_2 \\ &\quad -\frac{1}{2}R^{-1}[g - \hat{\psi}\hat{\varsigma}]^T \nabla \phi_2^T \tilde{W}_2\end{aligned} \tag{46}$$

so that $\lim_{t\to+\infty}\|\hat{u} - u^*\| = 0$ is true, which implies that the proposed control converges to its optimal solution.

2) In case when there are bounded approximation errors in the identifier NN and critic NN, then we know $\gamma \ne 0$. In this case, it can be shown that $\dot{V}$ is negative if

$$\|\tilde{W}_1\| > \sqrt{\gamma/a_1}, \quad \|\tilde{W}_2\| > \sqrt{\gamma/a_2}, \quad \|x\| > \sqrt{\gamma/a_3},$$
$$\|\upsilon_1\| > \sqrt{\gamma/a_4}, \quad \|\upsilon_2\| > \sqrt{\gamma/a_5}$$

which implies that the NN weight errors $\tilde{W}_1$, $\tilde{W}_2$ and the system states x are all UUB.

We finally evaluate the convergence property of the proposed optimal control. Recalling (46) with NN approximation errors ε_g and $\nabla \varepsilon_v$, we have

$$\begin{aligned}u - u^* &= -\frac{1}{2}R^{-1}[\hat{\psi}\hat{\varsigma}]^T \nabla \phi_2^T \hat{W}_2 + \frac{1}{2}R^{-1}g^T\left(\nabla \phi_2^T W_2 + \nabla \varepsilon_v\right) \\ &= \frac{1}{2}R^{-1}\left(g^T \nabla \phi_2^T \tilde{W}_2\right) + \frac{1}{2}R^{-1}[g - \hat{\psi}\hat{\varsigma}]^T \nabla \phi_2^T W_2 \\ &\quad -\frac{1}{2}R^{-1}[g - \hat{\psi}\hat{\varsigma}]^T \nabla \phi_2^T \tilde{W}_2 + \frac{1}{2}R^{-1}g^T \nabla \varepsilon_v\end{aligned} \tag{47}$$

which further implies the following fact

$$\begin{aligned}\lim_{t\to+\infty}\|\hat{u} - u^*\| \le \frac{1}{2}\lambda_{\max}(R^{-1})\Big[b_g\left(\phi_M\|\tilde{W}_2\| + \phi_\varepsilon\right) + \phi_M W_N\left(\|\tilde{W}_1\| + \|\varepsilon_g\|\right) \\ +\phi_M\|\tilde{W}_2\|\left(\|\tilde{W}_1\| + \|\varepsilon_g\|\right)\Big] \le \varepsilon_u\end{aligned} \tag{48}$$

where $\varepsilon_u > 0$ is a positive constant depending on the identifier NN and critic NN approximation errors. This completes the proof. □

Remark 5: As shown in the proof of Theorem 3, the terms $\Gamma x^T x$ concerning the system state, $\Upsilon_1 \upsilon_1^T \upsilon_1$ denoting the identifier error and $\Upsilon_2 \upsilon_2^T \upsilon_2$ defining the HJB residual error are all considered, such that the convergence of the system states to zero and the proposed control to its optimal solution can be proved in contrast to some available ADP schemes, where only UUB of the closed-loop system is achieved (Bhasin et al., 2013; Modares et al., 2013; X. Yang et al., 2014; Zhang, Cui et al., 2011).

4. Finite-time adaptation based optimal control

In this section, we will improve the design of adaptive laws to achieve finite-time convergence of NN weights.

We can design the adaptive laws for the identifier NN weights $\hat{W}_1$ and critic NN weights $\hat{W}_2$ as

$$\dot{\hat{W}}_1 = -\Gamma_1 \frac{P_1^T M_1}{\|M_1\|} \tag{49}$$

$$\dot{\hat{W}}_2 = -\Gamma_2 \frac{P_2^T M_2}{\|M_2\|} \tag{50}$$

where $\Gamma_1 > 0$, $\Gamma_2 > 0$ are constant learning gains.

Then we have the following Corollaries:

Corollary 1: For system (10) with adaptive law (49), if the regressor vector ϕ_1 is PE, then

i) for $\varepsilon_T = 0$, the NN error $\tilde{W}_1$ converges to zero in finite time.

ii) for $\varepsilon_T \neq 0$, the NN error $\tilde{W}_1$ converges to a compact set $\tilde{W}_1 = P_1^{-1}\upsilon_1$ in finite time.

Proof: We first analyze the derivative of $P_1^{-1}M_1$ with respect to time t. Consider the fact $M_1 = -P_1\tilde{W}_1 + \upsilon_1$ in (18), on can have that $P_1^{-1}M_1 = -\tilde{W}_1 + P_1^{-1}\upsilon_1$, such that

$$\frac{\partial P_1^{-1}M_1}{\partial t} = -\dot{\tilde{W}}_1 + \frac{\partial P_1^{-1}}{\partial t}\upsilon_1 + P_1^{-1}\dot{\upsilon}_1 = \dot{\hat{W}}_1 + \upsilon_1^* \tag{51}$$

where $\upsilon_1^* = -P_1^{-1}\dot{P}_1P_1^{-1}\upsilon_1 + P_1^{-1}\dot{\upsilon}_1$ denotes the lumped error due to the NN approximation errors $\varepsilon_f, \varepsilon_g$ and υ_1.

Now, consider the Lyapunov function as

$$V_{1f} = \frac{1}{2}M_1^T P_1^{-1}\Gamma_1^{-1}P_1^{-1}M_1 \tag{52}$$

Then it follows from (49)~(52) that

$$\begin{aligned}\dot{V}_{1f} &= M_1^T P_1^{-1}\Gamma_1^{-1}(\dot{\hat{W}}_1 + \upsilon_1^*) = -\frac{M_1^T P_1^{-1}P_1^T M_1}{\|M_1\|} + M_1^T P_1^{-1}\Gamma_1^{-1}\upsilon_1^* \\ &\le -\left(1 - \lambda_{\max}(\Gamma_1^{-1})\left\|P_1^{-1}\upsilon_1^*\right\|\right)\|M_1\|\end{aligned} \tag{53}$$

We now analyze the particular term υ_1^*. Consider the fact $\upsilon_1 = -\int_0^t e^{-\ell(t-r)}\phi_{1f}(r)\varepsilon_{Tf}^T(r)dr$, it can be verified that υ_1 and $\dot{\upsilon}_1$ remain bounded as long as ε_T and ϕ_1 are bounded. The matrices P_1 and $\dot{P}_1$ are also bounded for bounded ϕ_1. Moreover, the PE condition implies that P_1^{-1} is bounded in magnitude. Thus, by assuming bounded NN approximation errors $\varepsilon_f, \varepsilon_g$, the term υ_1^* is bounded. We can select sufficiently large adaptive gain Γ_1 such that $\lambda_{\max}(\Gamma_1^{-1}) < 1/\left\|P_1^{-1}\upsilon_1^*\right\|$ holds, then Eq.(53) can be reduced as $\dot{V}_{1f} \le -\mu_1\sqrt{V_{1f}}$, where $\mu_1 = \sqrt{2}\left(1 - \lambda_{\max}(\Gamma_1^{-1})\left\|P_1^{-1}\upsilon_1^*\right\|\right)/\lambda_{\max}(P_1^{-1})$ is a positive constant. In this case, from Lemma 1 (Bhat & Bernstein, 1998; Utkin, 1992) and (53), it follows that $\lim_{t\to\infty} V_{1f} = 0$ holds in finite time $t_c \le 2\sqrt{V_{1f}(0)}/\mu_1$, and thus $\lim_{t\to\infty} P_1^{-1}M_1 = 0$ is true in finite time t_c.

i) When there are no NN approximation errors, i.e., $\varepsilon_f = \varepsilon_g = 0$, such that $\varepsilon_T = 0$ and $\upsilon_1 = 0$, we know that $M_1 = -P_1\tilde{W}_1$ is true, and thus one can obtain that $\lim_{t\to\infty}\tilde{W}_1 = \lim_{t\to\infty}P_1^{-1}M_1 = 0$ holds in finite time t_c, i.e., the estimation error $\tilde{W}_1$ converges to zero in finite time. The convergence rate depends on the excitation level σ_1 and the learning gain Γ_1.

ii) In case when there are approximation errors $\varepsilon_f, \varepsilon_g$, we know that $M_1 = -P_1\tilde{W}_1 + \upsilon_1$. Then from the fact that $\lim_{t\to\infty} P_1^{-1}M_1 = 0$ holds in finite time t_c, one can conclude that the estimation error $\tilde{W}_1$ converges to the small set $\tilde{W}_1 = P_1^{-1}\upsilon_1$ in finite time.

Similar to Corollary 1, we have Corollary 2:

Corollary 2: For adaptive law (50) for critic NN with the regressor vector Ξ in (31) being PE, then

i) for $\varepsilon_{HJB} = 0$, the critic NN error $\tilde{W}_2$ converges to zero in finite time.

ii) for $\varepsilon_{HJB} \neq 0$, the critic NN error $\tilde{W}_2$ converges to a compact set $\tilde{W}_2 = P_2^{-1}\upsilon_2$ in finite time.

Proof: The proof of Corollary 2 is similar to that of Corollary 1, and thus is omitted here. □

We now summarize the main results of this section as:

Theorem 4: For system (1) with adaptive optimal control (29) and adaptive laws (49) and (50), if the regressor vectors ϕ_1 and Ξ are PE, then:

i) In the absence of approximation errors, the system states x and the NN errors $\tilde{W}_1$, $\tilde{W}_2$ converge to zero, and the adaptive control u in (29) converges to the ideal optimal solution u^* in (6), i.e., $u \to u^*$.

ii) In the presence of approximation errors, the system states x and the NN errors $\tilde{W}_1, \tilde{W}_2$ are UUB, and the adaptive control u in (29) converges to a small set around its ideal optimal solution u^* in (6), i.e., $\|u - u^*\| \le \varepsilon_u$ for a positive constant ε_u.

The proof of Theorem 4 can be conducted following the merit of the proof of Theorem 3, and we will not repeat it again. It should be noted that finite-time convergence of the proposed control is not claimed in Theorem 4 because the ideal optimal value function V^* should be included in the Lyapunov function, which creates extra complexity in the stability analysis. This issue will be further studied in our future work. However, the improved finite-time convergence of both identifier NN and critic NN can lead to better transient performance as indicated in Simulation.

5. Simulations

Consider a nonlinear CT affine system (Nevistic & Primbs, 1996) as

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix} + \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix} u \tag{54}$$

In the simulation, the system dynamics (including both the input dynamics $g(x)$ and drift dynamics $f(x)$) in (54) are assumed to be unknown; this is different to the ADP results where the input dynamics is assumed to be known (Bhasin et al., 2013). We reorganize system (54) in the form of (10) and use (14) to estimate the

unknown identifier weights $W_1=[\theta,\psi]=\begin{bmatrix}-1 & 1 & 0 & 0 & 0\\ -0.5 & 0 & -0.5 & 1 & 2\end{bmatrix}$ with $\phi_1(x,u)=\left[x_1,x_2,x_2(1-x_2(\cos(2x_1)+2)^2),u\cos(2x_1),u\right]^T$ being the regressor vector. The parameters used in the simulations are $k=0.001$, $\ell=6$, $\Gamma_1=400$. The initial identifier NN weights are $\hat{W}_1(0)=0$ and system states are $x_1(0)=3, x_2(0)=-1$. Fig. 2 shows the profile of no null elements of the estimated weights $\hat{W}_1$ with adaptive law (14), which all converge to their true values.

The proposed optimal control (29) is then evaluated. For this purpose, the matrices Q and R in the cost function (2) are chosen as identity matrices. Moreover, as shown in (Nevistic & Primbs, 1996; Vamvoudakis & Lewis, 2010), the optimal control (6) will be designed by choosing the optimal value function (4) as

$$V^*(x)=\frac{1}{2}x_1^2+x_2^2 \quad \text{and}$$

$$u^*=-\frac{1}{2}R^{-1}[g(x)]^T\frac{\partial V^*(x)}{\partial x}=-\left(\cos(2x_1)+2\right)x_2 \quad (55)$$

Similar to (Bhasin et al., 2013; Vamvoudakis & Lewis, 2010), we select the activation function for the critic NN as $\phi_2(x)=[x_1^2,x_1x_2,x_2^2]^T$, then the ideal critic NN weights $W_2=[0.5,0,1]^T$ are derived. The parameters for the critic NN learning are $\Gamma_2=200\text{diag}([0.3,1,1])$, $\ell_2=150$ and the initial critic NN weights are $\hat{W}_2(0)=[0\ 0\ 0]$. A critical issue in using the developed optimal control is to ensure the PE of the critic regressor vector. One standard approach is to introduce a dither signal $d(t)$ into the control signal first and remove it when the parameter convergence is retained. In this simulation, a small exploratory signal d(t)=0.8(sin^2(t)cos(t)+sin^2(2t)cos(0.1t) +sin^2(-1.2t)cos(0.5t)+sin^5(t)) is added for the first 1s as (Vamvoudakis & Lewis, 2010).

The estimated critic NN weights $\hat{W}_2$ with the adaptive law (34) are shown in Fig.3. One can find that the estimation $\hat{W}_2$ converges to a small set around the true values, i.e., $\hat{W}_2$=[0.5001,0.0005,1.0001]; this means that the designed adaptive optimal control (29) converges to its optimal control action in (55). It should also be noted that the novel update laws (14) and (34) based on the information of the parameter estimation errors lead to faster convergence of the NN weights as compared to (Vamvoudakis & Lewis, 2010). Moreover, the input dynamics $g(x)$ are unknown in this paper. Fig.4 shows the evolution of the system state, which can be stabilized by the suggested identifier-critic based optimal control. The required control action is provided in Fig.5. Finally, Fig. 6 shows the 3D plot of the error between the approximated value function $\hat{V}(x)=\hat{W}_2^T\phi_2(x)$ and the ideal value in (55), which is close to zero, i.e., good approximation of the value function is obtained.

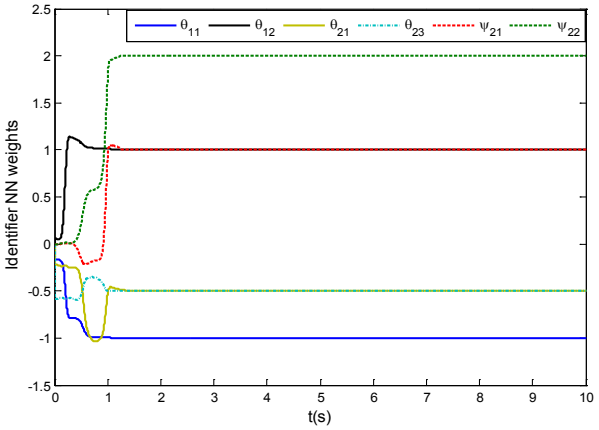

Fig. 2 Convergence of the identifier NN weights $\hat{W}_1$.

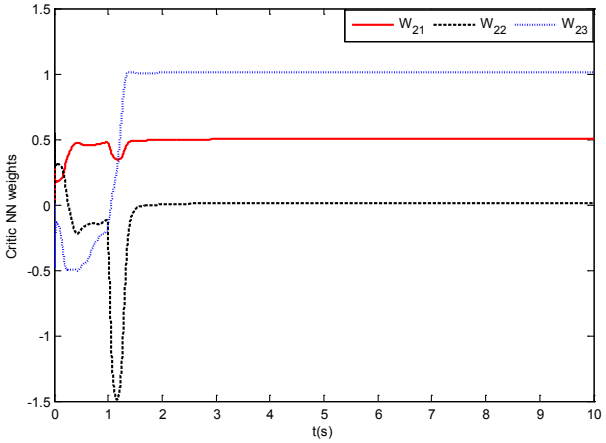

Fig. 3 Convergence of the critic NN weights $\hat{W}_2$.

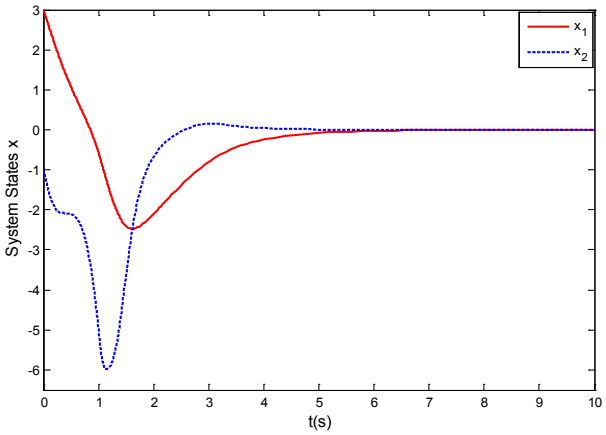

Fig. 4 Profile of system states.

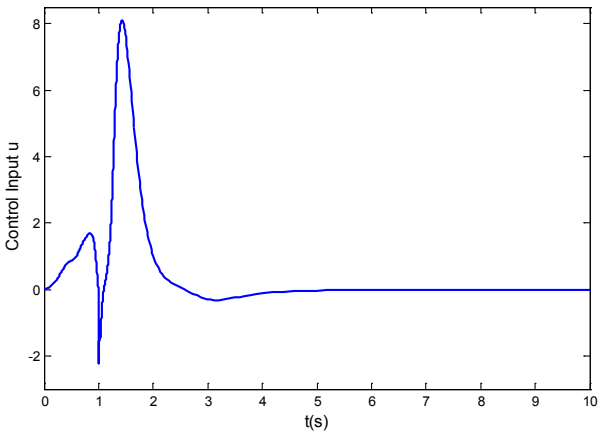

Fig. 5 The proposed control action.

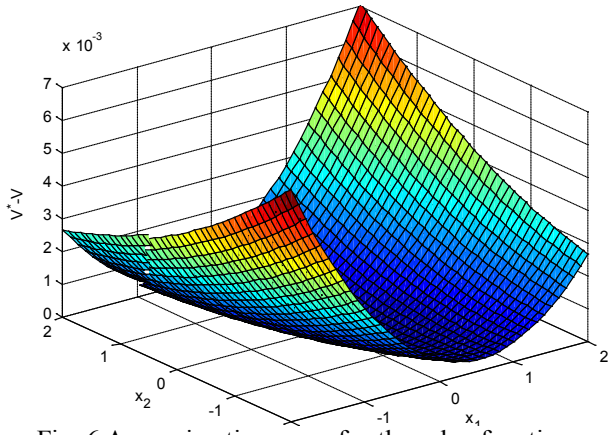

Fig. 6 Approximation error for the value function.

Finally, the proposed finite-time adaptive laws (49) and (50) are simulated, and the profiles of the identifier NN and critic NN weights are illustrated in Fig.7 and Fig.8, respectively. It is shown that the NN weights converge to a small set around their true values slightly faster than those of the adaptive schemes (14) and (34), i.e., the convergence time for the identifier NN with (14) and the critic NN with (34) are around 1s and 1.5s, while they are 0.5s and 1s for the finite-time approaches (49) and (50), respectively.

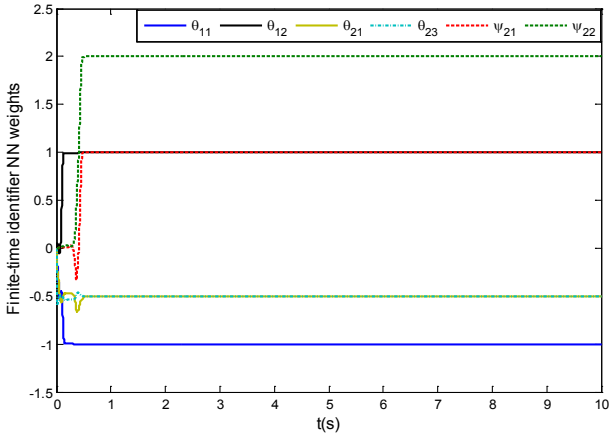

Fig. 7 FT convergence of identifier weights $\hat{W}_1$.

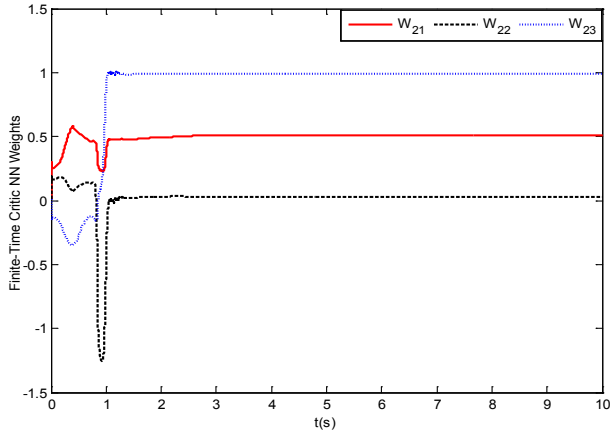

Fig. 8 FT convergence of the critic NN weights $\hat{W}_2$.

6. Conclusions

In this paper, an adaptive optimal control is developed for continuous-time affine nonlinear systems with completely unknown dynamics. An NN identifier was designed to estimate the unknown system dynamics, and a critic NN is used to online learn the solution of the HJB equation. The identifier NN and critic NN are then used to calculate the optimal control. This leads to a novel identifier-critic based ADP structure with a simplified dual NN approximation, where the actor NN is avoided. Novel adaptive laws based on the parameter estimation error are proposed to estimate the weights of both identifier NN and critic NN simultaneously. The proposed adaptations are further improved to achieve finite-time convergence. The effect of the identifier error on the control convergence is addressed and the stability of the closed-loop system is proved. Simulations are given to validate the efficacy of the proposed methods.

Acknowledgements

This work was supported by XXXXXX (No. XXX) and XXXX (No.XXX).

References

Abu-Khalaf, M., & Lewis, F. L. (2005). Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica, 41*(5), 779-791

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2008). Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38*(4), 943-949

Bellman, R. E. (1957). *Dynamic programming*: New Jersey:Princeton University Press.

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K. G., Lewis, F. L., & Dixon, W. E. (2013). A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica, 49*(1), 82-92

Bhat, S. P., & Bernstein, D. S. (1998). Continuous finite-time stabilization of the translational and rotational double integrators. *IEEE Transactions on Automatic Control, 43*(5), 678-682

Dierks, T., & Jagannathan, S. (2012). Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update. *IEEE Transactions on Neural Networks and Learning Systems, 23*(7), 1118-1129

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural computation, 12*(1), 219-245

Farza, M., Sboui, A., Cheerier, E., & M'Saad, M. (2010). High-gain observer for a class of time-delay nonlinear systems. *International Journal of Control, 83*(2), 273-280

Hanselmann, T., Noakes, L., & Zaknich, A. (2007). Continuous-time adaptive critics. *IEEE Transactions on Neural Networks, 18*(3), 631-647

Jiang, Y., & Jiang, Z.-P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica, 48*(10), 2699-2704

Jung, J. C., Huh, K., Lee, T.H. (2008). Observer design methodology for stochastic and deterministic robustness. *International Journal of Control, 81*(7), 1172-1182

Kamalapurkar, R., Walters, P., & Dixon, W. (2013). Concurrent learning-based approximate optimal regulation *52nd IEEE Conference on Decision and Control December* (pp. 6256-6261). Florence, Italy: IEEE.

Lewis, F. L., & Vrabie, D. (2009). Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine, 9*(3), 32-50

Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal control*: Wiley. com.

Liu, D., Huang, Y., Wang, D., & Wei, Q. (2013). Neural-network-observer-based optimal control for unknown nonlinear systems using adaptive dynamic programming. *International Journal of Control, 86*(9), 1554-1566

Liu, D., & Wei, Q. (2013). Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems. *IEEE Transactions on Cybernetics, 43*(2), 779-789

Liu, Y.-J., Tang, L., Tong, S., Chen, C. L., & Li, D.-J. (2015). Reinforcement Learning Design-Based Adaptive Tracking Control

With Less Learning Parameters for Nonlinear Discrete-Time MIMO Systems. *IEEE transactions on neural networks and learning systems, 26*(1), 165-176

Modares, H., Lewis, F. L., & Naghibi-Sistani, M. B. (2013). Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 24*(10), 1513 - 1525

Na, J., & Herrmann, G. (2014). Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuous-time unknown nonlinear systems. *IEEE/CAA Journal of Acta Automatica Sinica, 1*(4), 412-422

Na, J., Herrmann, G., Ren, X., Mahyuddin, M. N., & Barber, P. (2011). Robust adaptive finite-time parameter estimation and control of nonlinear systems *IEEE International Symposium on Intelligent Control (ISIC)* (pp. 1014-1019).

Na, J., Ren, X., & Zheng, D. (2013). Adaptive Control for Nonlinear Pure-Feedback Systems With High-Order Sliding Mode Observer. *IEEE Transactions on Neural Networks and Learning Systems, 24*(3), 370-382

Nevistic, V., & Primbs, J. A. (1996). *Constrained nonlinear optimal control: a converse HJB approach*. Technical Report.

Ni, Z., & He, H. (2013). Adaptive learning in tracking control based on the dual critic network design. *IEEE Transactions on Neural Networks and Learning Systems, 24*(6), 913-928

Qin, C., Zhang, H., & Luo, Y. (2014). Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming. *International Journal of Control, 87*(5), 1000-1009

Ren, X. M., Lewis, F. L., & Zhang, J. (2009). Neural network compensation control for mechanical systems with disturbances. *Automatica, 45*(5), 1221-1226

Sastry, S., & Bodson, M. (1989). *Adaptive control: stability, convergence and robustness*: Courier Dover Publications.

Si, J., Barto, A. G., Powell, W. B., & Wunsch, D. C. (2004). *Handbook of learning and approximate dynamic programming*: IEEE Press Los Alamitos.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*: Cambridge Univ Press.

Utkin, V. I. (1992). *Sliding modes in control and optimization*: Springer-Verlag Berlin.

Vamvoudakis, K. G., & Lewis, F. L. (2010). Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica, 46*(5), 878-888

Vrabie, D., & Lewis, F. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks, 22*(3), 237-246

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., & Lewis, F. L. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica, 45*(2), 477-484

Wang, D., Liu, D., Wei, Q., Zhao, D., & Jin, N. (2012). Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming. *Automatica, 48*(8), 1825-1832

Wang, F.-Y., Zhang, H., & Liu, D. (2009). Adaptive dynamic programming: an introduction. *IEEE Computational Intelligence Magazine, 4*(2), 39-47

Werbos, P. J. (1990). A menu of designs for reinforcement learning over time. *Neural networks for control*, 67-95

Werbos, P. J. (1992). Approximate dynamic programming for realtime control and neural modeling. In D. A. White & D. A. Sofge (Eds.), *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches* (pp. 67-95): New York: Van Nostrand Reinhold.

Xu, B., Yang, C., & Shi, Z. (2014). Reinforcement learning output feedback NN control using deterministic learning technique. *IEEE Transactions on Neural Networks and Learning Systems, 25*(3), 635-641

Yang, Q., & Jagannathan, S. (2012). Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 42*(2), 377-390

Yang, Q., Vance, J. B., & Jagannathan, S. (2008). Control of nonaffine nonlinear discrete-time systems using reinforcement-learning-based linearly parameterized neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 38*(4), 994-1001

Yang, X., Liu, D., & Wang, D. (2014). Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints. *International Journal of Control, 87*(3), 553-566

Zhang, H., Cui, L., Zhang, X., & Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks, 22*(12), 2226-2236

Zhang, H., Song, R., Wei, Q., & Zhang, T. (2011). Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming. *IEEE Transactions on Neural Networks, 22*(12), 1851-1862